实体及其属性的相关抽取技术

孟涛 曹雷 折闪电 程学旗

摘要： 信息抽取是当前搜索引擎与自然语言处理研究领域的核心技术之一，它用来对文本做匹配，以获得其中包含的各种实体以及它们的属性及关系。本文对实体及其属性的抽取做了简单介绍，包括基于规则的抽取技术和基于统计的抽取技术，并介绍了几个典型的系统实例，如：IE2、GATE和SystemT及它们的原理，最后简单介绍了我们在这个领域的工作成果。

关键词： 信息抽取，实体抽取，规则匹配

1 信息抽取技术简介

信息抽取是指从非结构化的信息源中抽取出特定的信息，并将抽取出来的信息以结构化形式进行保存（如保存到数据库或者 XML[1]文件中），供进一步查询和分析。抽取出来的信息类型包括实体、实体的关系以及实体的属性等。例如，从新闻报道中可以抽取出新闻事件相应的时间、地点、人物以及人物之间的关系。随着互联网信息数量的快速增长，信息抽取技术被大量用于分析网页和自由文本，包括在舆情监控（如匹配敏感信息）、电子商务（如抽取产品属性）、情感分析（如抽取褒贬特征）、甚至自然语言处理技术本身（如匹配词性标注或名词短语语法规则集）等诸多领域。对信息抽取技术的研究也已经持续了很多年，引文[1-4]对于信息抽取技术和系统进行了总结。

信息抽取的快速发展同 MUC，LRE，ACE，EMLIED 和 SIGHAN 等评测会议和项目计划密切相关。MUC(Message Understanding Conferences，讯息理解会议)由美国国防部高级研究计划署(DARPA)发起，目标是对不同的信息抽取系统就某一特定领域中的自由文本抽取的效果进行评测，从 1987 年至 1998 年共举办了 7 届。LRE(Linguistic Research and Engineering，语言研究与工程)由欧盟发起，资助项目为信息抽取开发工具和组件，包括从文档集中获取词典和抽取实体等。ACE(Automatic Content Extraction，自动内容抽取)是 MUC 停办后美国国家标准技术研究所(NIST)发起的自动内容抽取评测会议，主要目标是从文本中自动抽取特定信息，包括实体、实体关系和事件，从 1999 年开始举办至今；ACE 从 2003 年开始加入了针对汉语的评测。

SIGHAN 评测则关注中文处理，由计算语言学协会（Association for Computational Linguistics，ACL）的汉语处理特殊兴趣小组发起，2003 年首次举办，从 2006 年开始加入了对命名实体识别的评测。

一般来说，为完成一项信息抽取任务，需要事先提供一组规则直接定义抽取目标；或是提供一个标注文档集来对抽取范围作间接限定，信息抽取系统以此为基础从其它文档中寻找与用户定义相符的数据。因此，从抽取技术来说，一般可以分为基于规则的信息抽取技术和基于统计的信息抽取技术：前者是由用户提供抽取的规则，由抽取系统执行匹配，这种方法在特定领域一般具有很高的精确度；后者则是由用户提供标注的训练集，由抽取系统自动学习抽取的准则，这种方法对文本中的噪音具有较好的健壮性，有相对较高的召回率。

[1] Extensible Markup Language，可扩展标记语言

下面将从基于规则的抽取方法、基于统计的抽取方法、系统实例以及我们的工作等几个方面分别进行介绍。

2 基于规则的实体抽取方法

常见实体包括人物、地点、机构、日期等。基于规则的抽取方法利用规则进行实体抽取。规则可以由人工编写，也可以通过自动学习的方法生成。这种方法的优点是规则表现形式简单，易于被人所理解，并且便于维护和扩展。规则方法所使用的抽取规则之间彼此可以相互独立；也可以存在依赖关系，例如组成一个上下文无关文法。

2.1 规则的表现形式

用户信息抽取的规则一般由一组条件和一个对应的动作组成。条件包括命名实体特征和上下文环境特征；动作指将命名实体标记为相应的类别。规则使用的实体特征可以包括：字符串、词形特征（例如大小写和标点符号等）、词性特征、词所属的概念类别以及上下文环境中词的共现等。当输入文本满足规则规定的条件时，相应的动作就被触发。

用来进行实体抽取的规则按照表现形式可以分为三类：

1. 定义实体本身：这类规则同时限定实体的组成形式以及它的上下文边界。例如定义英文中的公司名时，可以限制实体的后缀在形式上首字母为大写，且它必须出现在一个由 LLC、Corp.等组成的领域词典中。
2. 定义实体的边界：有时无法对实体的组成形式给出准确而全面的定义，但可以给出实体的上下文边界的定义。例如定义 journal 这个实体的前方边界为“to appear in”。
3. 定义多个实体：可以利用实体处于同一个上下文环境来做语义消歧，进而同时定义多个实体，例如 Number 在和 Bedrooms 在一起时可以定义为房间数或是租金。

在实际的规则系统中，经常出现大量规则相互冲突的情况，譬如规则匹配到的实体的文本相互重叠。为解决这个问题，有两种策略：

- 将规则集合看成是彼此无序。规则之间相互独立，如果发生冲突，则按照某种预定的策略来解决，比如选择匹配区域较长的文本。
- 将规则集合看成是有序的，匹配前先定义好规则的顺序，然后按照顺序执行规则。这种序列关系可以以规则覆盖率或者准确率作为排序指标。

2.2 规则库构建

可以通过人工编写和自动学习两种方式来产生实体抽取的规则。

基于人工编写的方式，就是领域专家或者语言学家手工编写抽取规则。该方法需要规则的编写者具备丰富的领域知识和语言学知识，同时还需要大量的人工分析来进行总结归纳，是一项非常耗费时间和人力的工作，但是这种方法也往往能够得到非常高的准确率。

基于自动学习的方式，是指在标记好的语料库上进行训练，得到规则。该方法不但能节省人力和时间，还可以挖掘出人工观测所不易发现的特征；但是该方法需要大量标注好的训练语料，如果训练语料不充分或者标注质量不高，会导致训练效果较差；另外，训练语料的标注本身也是一个耗时耗力的工作。

2.3 规则自动学习

自动学习的输入是一个已经标注出待抽取结果的文档集，输出是一组用于抽取的规则。已有的规则自动学习方法有序列覆盖方法(Sequential Covering Algorithm)和基于转换的学习方法(Transformation based Learning)两类，它们分别对应于规则从无到有的建立和对已有规则的选取两步。

1. **序列覆盖方法**

序列覆盖算法是一种归纳学习算法，包括自底向上(Bottom-Up)和自顶向下(Top-Down)两种。自底向上的方法是一个泛化的过程，代表性的算法是(LP)2[5]，基本流程如下：

（1） 训练集中选择某个样本实例，将其所对应的最具体的规则作为种子规则。
（2） 泛化这条种子规则（放松规则中的某个条件限制或者移除该条件），直到符合某种标准，将泛化后的规则加入到规则集中。
（3） 将泛化后的规则所覆盖到的样本从训练集中移走。
（4） 重复上述步骤直到训练集合为空。

而自顶向下的方法是一个具体化的过程，代表性的算法包括 FOIL[6-7]，基本流程如下：

（1） 从训练集中选择某个样本实例，将其所对应的具体的规则作为种子规则。
（2） 首先将种子规则的条件移除，然后不断加入条件限制，对该规则进行具体化，直到符合某种标准，将该规则加入目标规则集合。
（3） 将规则所覆盖到的样本从训练集中移走。
（4） 重复上述步骤直到训练集合为空。

从上述两种算法流程可以看出，由于搜索方向不同，自底向上的方式倾向于产生比较具体的规则，训练过程中不断提高所产生规则的召回率；而自顶向下更倾向于产生比较泛化的规则，训练过程中不断提高所产生规则的精度。

2. **基于转换的学习方法**

基于转换的学习方法用来学习一组有序的规则，最早由埃里克. 布瑞尔（Eric Brill）提出运用于英文词性标注[8]。该算法过程非常简洁，且能得到不错的效果，流程如下：

输入：训练集合，规则模板集
输出：一组有序的规则
（1） 为训练集合的样本分配一个初始的类别标号。
（2） 遍历训练集合，如果某个训练样本被分配的类别标号不同于其真实类别，则根据规则模板，生成相应的规则，每条规则计算一个评分。一般情况下，评分取值为该规则将错误类别标号改成正确类别标号的样本数与该规则将正确类别标号改成错误类别标号的样本数之差。
（3） 选择一条评分最高的规则。
（4） 如果该规则的评分大于阈值，则返回到步骤（2），否则结束执行。

算法流程中的第一步中要为样本分配初始的类别标号。这个分配方法可以视具体应用而定。例如，可以利用成熟的分类方法先为该样本分配一个最有可能的类别标号，在此基础上，再进行转换处理。所以，基于转换的学习多是作为后续处理过程，来提高此前标记过程的准确度。

2.4 规则执行的性能优化

基于规则的抽取系统一般采取有穷状态自动机（Finite-State Transducer）技术来优化规则的匹配。多数规则，包括词典和正则表达式以及它们的集合，都可以用自动机来表示。更进一步，实体类别间的关系通常可以由一个上下文无关文法来表示，这使得我们可以用层叠有穷状态自动机（Cascaded Finite-State Transducers）技术来表示一个规则集并将它与输入文本进行高效的匹配。这种传统的技术路线在当前受到了两方面的挑战：一方面，抽取目标更加复杂，用户往往需要在特定的知识领域做抽取，这要求抽取系统能在不同的领域让用户自己声明抽取目标后针对用户需求做抽取；另一方面，抽取系统针对的数据集规模太大（通常是数以亿计的文档），这就要求抽取系统必须采取一定的优化策略来提高数据处理效率。近两年提出的 SystemT[9-10]和 DBLife[11]就是属于这一类试图做高效率声明式信息抽取的系统。它们使用通用的信息抽取语言(AQL 和 Datalog)来声明抽取的实体目标，将规则语句用一定的编译技术进行优化，并在处理海量数据时针对数据特性动态调整规则的执行顺序以避免不必要的匹配，从而提高速度，因此具有较好的效果。

3 基于统计学习的实体抽取方法

统计方法一般将文本切分成若干片段之后对每个片段来进行自动分类，选出那些构成实体及属性的片段输出。切分可能存在不同的层次，可以切分成字或切分成词，这使得分类的特征有所不同。基于这些切分后的文本，首先人工选择一部分标注出那些实际构成实体的片段，形成标注语料库；然后根据语料库，利用统计方法，训练出根据文本片段的上下文环境来推测它构成实体及属性的概率模型；最后将模型应用于那些未标注的文本进行分析，选择可能性最大的输出。尽管这种方法得到的模型往往不为人所理解，但它对文本中的噪声具有较强的健壮性。

3.1 特征选择及其表示

文本片段的标注类别取决于切分的层次和实体的种类。如果切分到词和短语级别，则每个片段的类别直接与实体的种类相对应。如果切分到字的级别，即单个文本片段不足以构成实体，此时有两种不同的分类方法：假定某类实体为 A，则文本片段的类别在 BCEO 分类方法下有 A_Begin, A_Continue, A_End 和 Other 四类，对应着 A 类实体文本的开始、中间、结束和与之无关；在 BIO 分类方法下有 A_Begin, A_Inside 和 Other 三类，分别对应于开始、在其中和与之无关。因此，当实体集合中有 $\boldsymbol{n}$ 类实体时，在 BCEO 方法下，每一个文本片段可能属于 $3\times\boldsymbol{n}+1$ 种不同类别；在 BIO 方法下，每一个文本片段可能属于 $2\times\boldsymbol{n}+1$ 种不同类别。

用于对文本片段进行分类的特征可以统一表示为：

$$\boldsymbol{f}:(\boldsymbol{x},\boldsymbol{y},\boldsymbol{i})\to\mathbb{R}$$

这里假定文本被切分为一个序列 $\boldsymbol{x}=(\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_n)$，函数 $\boldsymbol{f}$ 表示文本为 $\boldsymbol{x}$ 时将第 $\boldsymbol{i}$ 个词 $\boldsymbol{x_i}$ 标记为 $\boldsymbol{y}$ 的情况。例如：

$f_1(x,y,i)=\{x_i="毛", x_{i+1,i+2}="泽东"\}\bullet\{y=Person_Start\}$ 表示在将文本切分到字时，当 x_i 是“毛”、 x_{i+1} 是“泽”、 x_{i+2} 是“东”时，将 x_i 标记为 $Person_Start$ 的特征；

$f_2(x,y,i)=\{x_i\in Person_dictionary\}\bullet\{y=Person\}$ 表示在将文本切分到词和短语时，当 x_i 属于某个词典 $Person_dictionary$ 时， x_i 标记为 $Person$ 的特征。

特征的选择可以来源于多个角度。例如，文本片段本身、文本片段属于某个实体对应的特别的词典、文本片段在字形上具有的某种特征（英文字符大小写、数字字符）、文本片段的上下文环境中的其它字词的特征，等等。

这样，通过将文本进行切分后对每个片段赋予对应的实体类别，选择片段属性及其上下文作为分类特征，我们可以将实体抽取问题转化成对文本切分之后对切分得到的序列进行分类的问题，即计算 $Pr\{y|x\}$。

3.2 标注模型

如果忽略实体类别间的相互依赖关系，即判断文本片段属于某类别仅取决于它的文本属性与上下文特征，可以简单地将实体抽取看成普通的对文本片段进行分类的问题。例如，使用最大熵算法(ME)[12]和支持向量机算法(SVM)[13]进行分类。它们根据输入文本中的字（词或短语）的属性和上下文环境特征来对该文本片段标记类别。但更多时候，相邻的文本片段的实体类别存在一定的依赖关系，因此需要从标注序列整体来进行分类，无法将各个文本片段独立处理。这种思路的分类方法包括 HMM[2][14]，MEMM[3][15]和 CRF[4][16]等，它们可以同时考虑到文本特征和类别间的依赖。

HMM，MEMM 和 CRF 常用来为序列标注建模，例如引文[17]利用采用基于角色的 HMM 模型对中国人名进行抽取；引文[18]结合 MEMM 模型与规则方法做实体抽取；引文[19]用 CRF 模型对命名实体做抽取。HMM 的一个主要缺陷是其输出独立性假设限制了特征的选择，导致在一个模型中不能考虑多个特征；针对该问题，引文[20-21]分别在传统的 HMM 基础上进行了改进，并将新的模型分别应用到了中文和英文命名实体抽取上。MEMM 相比于 HMM 不存在特征受限的问题，但它在每一节点都要进行归一化，只能找到局部最优值，同时也带来了标记偏置的问题。CRF 则可以任意选择特征，它继承了 MEMM 的优点，同时又解决了 MEMM 的标记偏置问题。它并不在每一个节点进行归一化，而是所有特征进行全局归一化，由此可以求得全局的最优值。因而一般认为 CRF 更适合于解决序列标注问题。

根据 CRF 的马尔可夫性，文本片段序列中的每一个被标记元素只依赖于与它相邻的片段的类别，因此相邻元素之间的类别的依赖性可在以下形式中体现：

$$\psi(y_{i-1},y_i,x,i)=e^{\sum_{k=1}^{K} w_k f_k(y_i,x,i,y_{i-1})}=e^{w\bullet f(y_i,x,i,y_{i-1})}$$

它对应在第 $i-1$ 和 i 个位置分别标记为类别 y_{i-1} 和 y_i 的概率。此时一个文本片段序列 x 被标记为 y 的条件概率根据哈默斯里-克利福德（Hammersley-Clifford）定理可以表示为：

$$Pr(y|x,w)=\frac{1}{Z(x)}\prod_{i=1}^{n}\psi(y_{i-1},y_i,x,i)=\frac{1}{\sum_y e^{w\cdot f(x,y)}}e^{\sum_{i=1}^{n} w\cdot f(y_i,x,i,y_{i-1})}$$

其中， $Z(x)$ 是一个归一化因子，且 $Z(x)=\sum_y e^{W\cdot f(x,y)}=\sum_y e^{W\cdot\sum_{i=1}^{n} f(y_i,x,i,y_{i-1})}$ 。

[2] Hidden Markov Models，隐马尔科夫模型
[3] Maximum Entropy Markov Models，最大熵马尔科夫模型
[4] Conditional Random Field，条件随机场

3.3 训练和分类

基于上述 $Pr(y|x,w)$ 模型，若将预先标出实体的训练集记为 $D=\{(x_l,y_l)\}_{l=1}^{N}$，可以根据它来估计其中的模型的参数 w。用到的训练算法通常有两类：最大似然估计和最大边界估计。

以最大似然估计为例来说明训练的基本过程。似然函数可以表示为：

$$L(w)=\sum_l log\,Pr(y_l|x_l,w)=\sum_l(w\cdot f(x_l,y_l)-log\,Z_w(x_l))$$

为避免过拟合导致部分参数偏差太大，可以在似然函数中加入一项 $-\|w\|^2/C$。因而训练目标可以表示为：

$$\max_w(L)=\max_w\sum_l(w\cdot f(x_l,y_l)-log\,Z_w(x_l))-\|w\|^2/C$$

上式是 L 的凸函数，因此，当 L 取最大值时，对 w 的梯度为 0，即：

$$\begin{aligned}\nabla L(w)&=\sum_l f(x_l,y_l)-\frac{\sum_{y'} f(y',x_l)e^{w\cdot f(x_l,y')}}{Z_w(x_l)}-2w/C\\&=\sum_l f(x_l,y_l)-E_{Pr(y'|w,x_l)}f(x_l,y')-2w/C\\&=0\end{aligned}$$

基于该式，可以使用迭代的方式来计算 w，相关的迭代求解 w 的算法包括半牛顿法[22]和随机梯度法[23]等。假定序列长度为 n，类别种类为 m，则序列一共有 $O(m^n)$ 种可能的标记。这种指数复杂性在训练过程和分类过程中都可以通过动态规划方法进行优化，将其复杂性降到 $O(nm^2)$；例如，对 $E_{Pr(y'|w,x_l)}f(x_l,y')$ 和 $Z_w(x_l)$ 都可以通过动态规划方法快速计算。

4 抽取系统介绍

一般来说，信息抽取系统可以定义为一个级联的转化器或者功能模块，使用手工或自动生成的规则为输入文本添加结构信息或去掉不相关信息。添加的结构信息可能来源于句法分析、词法分析、语法分析、语义分析和它们的级联等。本章将介绍几个有代表性的抽取系统，包括 IE2、ANNIE 和 SystemT。

4.1 IE2

IE2 [24]是早期有代表性的抽取系统之一。它在 MUC-7 中取得最好的效果。IE2 主要基于手工编写的规则进行抽取，由六个功能模块组成，如下图所示：

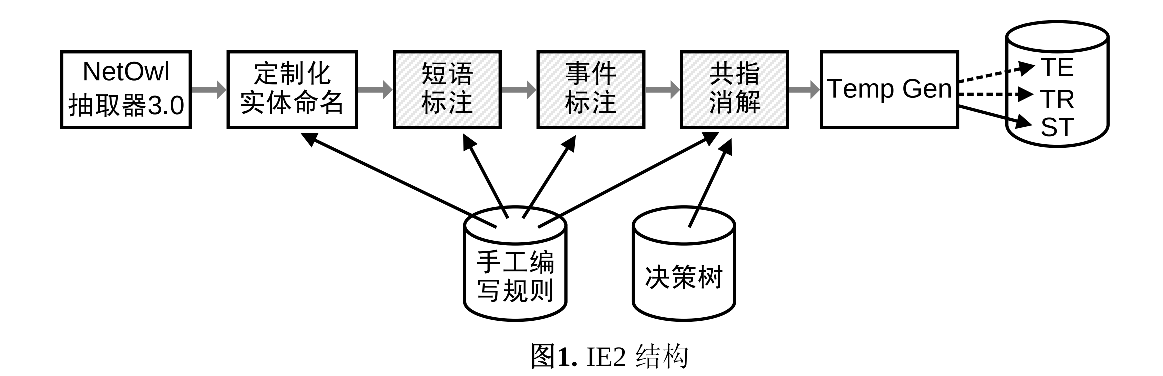


图1. IE2 结构

第一个模块(NetOwl)用来识别一般的通用命名实体，包括时间、地点、人名和数字等；第二个模块用来识别与领域相关的命名实体，如交通事故领域的命名实体包括飞机，轿车或者轮船等，得到的实体用 SGML 标记；短语标注模块（PhraseTag）基于前面结果发现更加复杂的名词短语及其包含的命名实体；事件标注模块（EventTag）识别句子中的事件或者事件片段；共指消除模块（Discourse Module）把指向同一实体的名词短语合并；最后一个模块 TempGen 合并隶属于同一个事件的事件片段，并按照规定的格式输出事件。

4.2 ANNIE

ANNIE 是 GATE[25]的核心组件之一，提供实体及属性的抽取服务，由英国谢菲尔德（Sheffield）大学开发，并公开了 Java 源代码。GATE 及 ANNIE 从 1995 年开始开发至今，支持多种语言和不同领域的信息抽取等自然语言处理技术，在学术界和工业界有着重要的影响。图 2 是 ANNIE 及其早期系统 LaSIE[26]的结构：

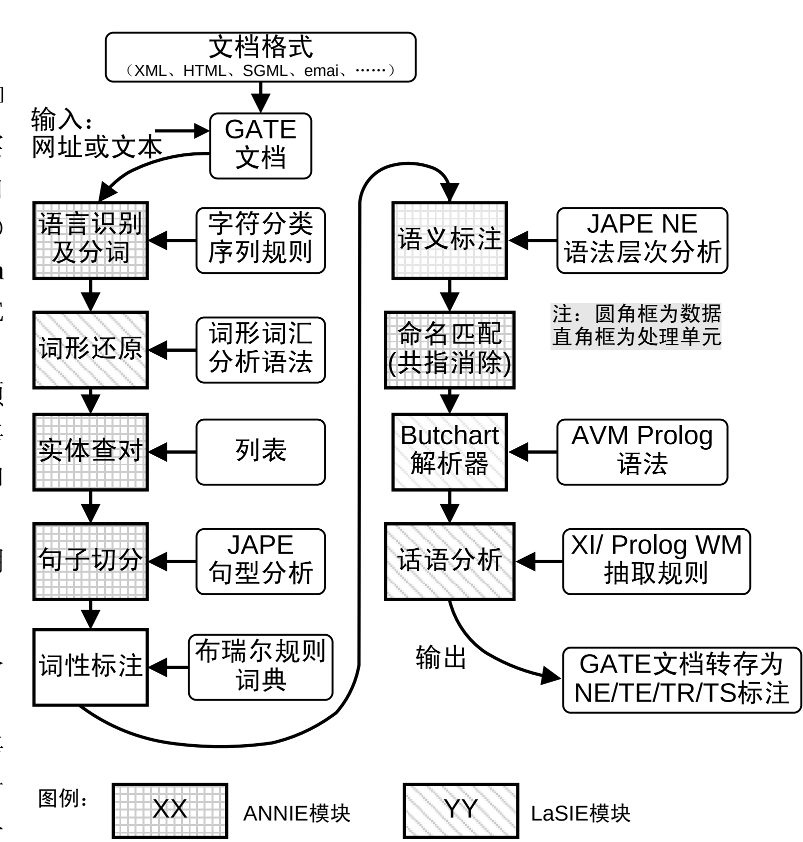


图2. ANNIE 及其早期系统 LaSIE 的结构

其中，语种识别及分词（Unicode Tokeniser）用于识别输入文本的语言并进行分词；实体查对（Gazetteer lookup）用于对词作标注，将其与用户事先为实体定义的词典建立联系，使用有穷状态自动机来加快匹配；句子切分（Sentence Splitter）用于分句；语义标注（Semantic Tagger） 和命名匹配（Name Matcher）是最重要的部分，它实现了一个通用的信息抽取语言 JAPE，让用户根据语法和上下文环境以将 JAPE 代码嵌在 Java 代码内的方式编程自定义抽取目标，因而具有广泛的适用性。

4.3 SystemT

SystemT[9]由 IBM 阿玛丹（Almaden）研究中心近两年基于稍早的 Avatar[27]抽取系统研究而成。它的一个主要创新在于，提出了一套从 SQL 扩展的信息抽取语言，与 DBLife 系统使用的 Datalog[11]语言和 ANNIE 使用的 JAPE 语言相似，能够对任意实体及其关系做声明式抽取；同时它对规则的执行顺序做了一定的优化，以适应对海量网络数据的抽取匹配带来的效率问题。其结构如图 3 所示。在该图中，用户使用 AQL 编写的实体抽取语句被优化器（Optimizer）编译成内部形式，然后被提交给运行时环境对输入文本流进行匹配以寻找实体。运行环境每次在内存中处理一个文档以避免不必要的磁盘读写开销。主要优化策略包括：（1）规则重写，即对正则表达式等效率较低的规则用更简单的字符串规则进行重写，以及将多个实体的字符串规则集在内存中同时匹配（Shared Dictionary Matching），这样可以大幅减小在 CPU 和磁盘读写上的时间代价；（2）规则内部优化，对复杂的诸如 JOIN 之类的操作符，分析规则内部不同子句之间的依赖性，调整它们的执行顺序，以避免不必要的子句匹配开销。

例如在布尔规则 AND 中，左子句匹配失效则避免继续对右子句进行匹配。

5 我们的工作

5.1 技术挑战

我们研发的实体及属性抽取系统主要应用于网络舆情分析。实际环境中待抽取的网页常常数以亿计，而关注的目标往往存在于不同的领域，包括人物、机构、地点等简单的实体抽取，也包括人物属性和人物关系等复杂的抽取任务。这使得信息抽取在工程中必须重点解决抽取系统的效率和抽取技术的通用性问题。针对这些难题，我们的技术路线是：首先建立概念及属性的描述语言，通过编写该语言的代码定制抽取目标，或通过标注语料库自动训练生成该语言的代码来制定抽取目标，最后将输入文本与由用户代码编译而成的自动机相匹配获得实体及属性。

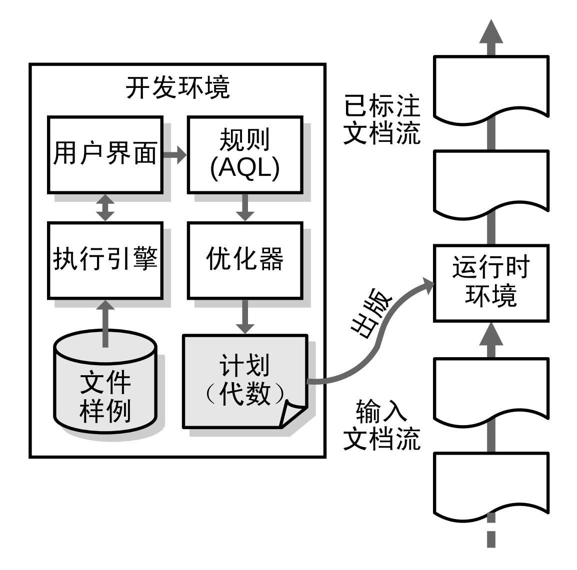


图3. SystemT 的结构

5.2 我们的工作

在信息抽取方面，我们结合实际项目需求，并跟踪国际前沿，开展了多项研发工作。这些工作在实际工程应用中发挥了重要的作用，包括：

高效的实体及属性抽取系统：该抽取系统由一组规则模板组成，用户可以选择模板和它们的组合来制定目标实体的抽取方式；每个规则在描述中体现为一个谓词。用户可以使用该系统在他的领域内编写代码完成对复杂实体的抽取工作。

抽取规则学习系统：用户可以提供标注语料库，并配置一些简单的模板，然后系统会根据统计方法自动训练出最适合该标注语料库的抽取模板，使得用户能避免过多参与规则的编写和配置。

针对人物属性与关系的抽取技术：针对舆情分析的实际项目需求，我们开发了一套基于上述系统的对人物属性和关系进行抽取的技术，它可以从文本集合中自动抽取人物实体、属性及人物关系。

6 结论与未来工作

信息抽取是当前搜索引擎与自然语言处理研究领域的核心技术之一，也是舆情分析的重要工程基础。尽管信息抽取已经发展了二十年，但在运行效率和通用性等方面仍然存在挑战。本文介绍了实体及属性的抽取技术，并总结了我们在这个领域的工作成果。未来工作中，我们一方面会根据现有基础，结合 SystemT 和 DBLife 等领先技术的特长对我们的系统在方法上和理论上进行改进；另一方面，会将它部署到更多的舆情数据分析的环节中。

参考文献：

[1] Sarawagi, S., *Information Extraction*. Found. Trends databases, 2008. 1(3): p. 261-377.

[2] Turmo, J., et al., *Adaptive information extraction.* ACM Comput. Surv., 2006. 38(2): p. 4.

[3] 李保利, 陈., 俞士汶, *信息抽取研究综述.* 计算机工程与应用, 2003. 39(10).

[4] 赵军, *命名实体识别, 排歧与跨语言关联.* 中文信息学报, 2009. 23(2).

[5] Ciravegna, F., *Adaptive Information Extraction from Text by Rule Induction and Generalisation, in the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI 2001*, B. Nebel, Editor. 2001, Morgan Kaufmann: Seattle, Washington. p. 1251-1256.

[6] Mitchell, T.M., *Machine Learning*. 1997: McGraw-Hill, Inc. 432.

[7] Quinlan, J.R., *Learning Logical Definitions from Relations*. Mach. Learn., 1990. 5(3): p. 239-266.

[8] Brill, E., *Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging*. Comput. Linguist., 1995. 21(4): p. 543-565.

[9] Krishnamurthy, R., et al., *SystemT: a system for declarative information extraction*. SIGMOD Rec., 2008. 37(4): p. 7-13.

[10] Reiss, F., et al., *An Algebraic Approach to Rule-Based Information Extraction, in Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*. 2008, IEEE Computer Society. p. 933-942.

[11] Shen, W., et al., *Declarative information extraction using datalog with embedded extraction predicates, in Proceedings of the 33rd international conference on Very large data bases*. 2007, VLDB Endowment: Vienna, Austria. p. 1033-1044.

[12] Chieu, H.L. and H.T. Ng, *A maximum entropy approach to information extraction from semi-structured and free text, in Eighteenth national conference on Artificial intelligence.* 2002, American Association for Artificial Intelligence: Edmonton, Alberta, Canada. p. 786-791.

[13] Sun, A., et al., *Using support vector machines for terrorism information extraction, in Proceedings of the 1st NSF/NIJ conference on Intelligence and security informatics*. 2003, Springer-Verlag: Tucson, AZ, USA. p. 1-12.

[14] Rabiner, L.R., *A tutorial on hidden Markov models and selected applications in speech recognition, in Readings in speech recognition*. 1990, Morgan Kaufmann Publishers Inc. p. 267-296.

[15] McCallum, A., D. Freitag, and F.C.N. Pereira, *Maximum Entropy Markov Models for Information Extraction and Segmentation, in Proceedings of the Seventeenth International Conference on Machine Learning*. 2000, Morgan Kaufmann Publishers Inc. p. 591-598.

[16] John Lafferty, A.M.a.F.P., *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, in Proc. 18th International Conf. on Machine Learning.* 2001, Morgan Kaufmann, San Francisco, CA. p. 282-289.

[17] 张华平, 刘群, *基于角色标注的中国人名自动识别研究.* 计算机学报, 2004. 27(1).

[18] Fresko, M., B. Rosenfeld, and R. Feldman, *A hybrid approach to NER by MEMM and manual rules, in Proceedings of the 14th ACM international conference on Information and knowledge management.* 2005, ACM: Bremen, Germany. p. 361-362.

[19] McCallum, A. and W. Li, *Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, in Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4.* 2003, Association for Computational Linguistics: Edmonton, Canada. p. 188-191.

[20] Fu, G. and K.-K. Luke, *Chinese named entity recognition using lexicalized HMMs*. SIGKDD Explor. Newsl., 2005. 7(1): p. 19-25.

（下转第 6 页）